



Міністерство освіти і науки України
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"
Фізико-технічний інститут

“КРИПТОГРАФІЯ”
Комп’ютерний практикум
Робота № 1

Виконав:

студент групи ФБ-93 Проценко О.А.

Перевірила:

Селюх П. В.

1. Назва роботи

Експериментальна оцінка ентропії на символ джерела відкритого тексту.

Мета

Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

2. Особливості роботи:

- Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку $1H$ та $2H$ за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення $1H$ та $2H$ на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення $1H$ та $2H$ на тому ж тексті, в якому вилучено всі пробіли.
- За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.
- Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

3. Робота виконана 06.10.2021 – 12.10.2021 року в місті Київ. Робота створена на мові програмування Python.

4. Робота над завданням

- Програма, яка підраховує частоти букв та біграм (що перетинаються і не перетинаються). Для визначення кількості букв (пробілу) використав звичайну посимвольну акумуляцію. Для визначення кількості біграм використав контейнер “collections” а саме об'єкт “Counter” в парі з функцією “range(start, stop, step)” де і зазначав крок дій. В результаті дані використовую для підрахунку їх частотності.

Складність полягала в недостатчі знань роботи в мові програмування. Все було вирішено, прочитавши потрібні інструкції до команд та саму документацію щодо лабораторної.

Частота букв без пробілу

```
o >>> 0.11275224397101763 time(s)
e >>> 0.08453768789877798 time(s)
a >>> 0.07721855736995782 time(s)
и >>> 0.06976316643235644 time(s)
н >>> 0.0640250892181248 time(s)
т >>> 0.060097328863415164 time(s)
с >>> 0.050762409430085434 time(s)
в >>> 0.04583324321401536 time(s)
л >>> 0.04576186871417757 time(s)
р >>> 0.042398615767275875 time(s)
к >>> 0.04061641613496269 time(s)
д >>> 0.03115388774737753 time(s)
м >>> 0.03007245593165351 time(s)
у >>> 0.029680977614361413 time(s)
п >>> 0.028376770844598248 time(s)
ь >>> 0.019610684546339352 time(s)
```

Ы >>> 0.018968314047799285 time(s)
 Я >>> 0.018596301503190225 time(s)
 Ч >>> 0.018496809776143614 time(s)
 Ъ >>> 0.01831296636747053 time(s)
 Г >>> 0.017633827187195847 time(s)
 З >>> 0.01697847950686709 time(s)
 Ж >>> 0.010822969611765978 time(s)
 Й >>> 0.010444468476262571 time(s)
 Ш >>> 0.009771817886882233 time(s)
 Х >>> 0.009546880069211637 time(s)
 Ю >>> 0.006330701849248405 time(s)
 Ц >>> 0.003611982264518222 time(s)
 Щ >>> 0.0031664323564399265 time(s)
 Э >>> 0.0027100681302043906 time(s)
 Ф >>> 0.0015377960419595544 time(s)
 Ь >>> 0.0003071266356656213 time(s)
 Ё >>> 0.00010165459067805775 time(s)

Ентропія букв без пробілу >>> 4.472830206355587

$$R = 1 - \frac{4.472}{5} = 0.1056 \quad 10.56\%$$

Частота букв з пробілом

_ >>> 0.1611100527084524 time(s)
 О >>> 0.0945867240018507 time(s)
 Е >>> 0.07091781654555517 time(s)
 А >>> 0.06477787152201327 time(s)
 И >>> 0.058523619011330956 time(s)
 Н >>> 0.053710003719529345 time(s)
 Т >>> 0.05041504504259315 time(s)
 С >>> 0.042584074971196324 time(s)
 В >>> 0.03844904698400602 time(s)
 Л >>> 0.038389171633599145 time(s)
 Р >>> 0.03556777254624464 time(s)
 К >>> 0.03407270319063042 time(s)
 Д >>> 0.026134683250324326 time(s)
 М >>> 0.0252274809714322 time(s)
 У >>> 0.024899073746473252 time(s)
 П >>> 0.02380498779812935 time(s)
 Ь >>> 0.016451206125429788 time(s)
 Ы >>> 0.015912327971767865 time(s)
 Я >>> 0.015600250387828975 time(s)
 Ч >>> 0.015516787778170898 time(s)
 Ъ >>> 0.015362563390759238 time(s)
 Г >>> 0.014792840359614983 time(s)
 З >>> 0.014243075778606356 time(s)
 Ж >>> 0.009079280407152383 time(s)
 Й >>> 0.00876175960954014 time(s)

ш >>> 0.008197479792069237 time(s)
 х >>> 0.008008781718059675 time(s)
 ю >>> 0.005310762140634497 time(s)
 ц >>> 0.003030055611499696 time(s)
 щ >>> 0.0026562882725961407 time(s)
 э >>> 0.002273448910903664 time(s)
 ф >>> 0.0012900416405846012 time(s)
 ь >>> 0.0002576454472053634 time(s)
 ё >>> 8.527701421585971e-05 time(s)

Ентропія букв з пробілом >>> 4.389170847948409

$$R = 1 - \frac{4.389}{5} = 0.1222 \quad 12.22\%$$

Біграма з пробілом 1 крок (перші 20)

о_ >>> 0.021331993090734906 time(s)
 и_ >>> 0.01835273540127444 time(s)
 е_ >>> 0.018309189612877943 time(s)
 _п >>> 0.017079021090676848 time(s)
 _н >>> 0.016237135848344533 time(s)
 _в >>> 0.015230139491675497 time(s)
 _с >>> 0.015226510675975788 time(s)
 а_ >>> 0.014711218846617217 time(s)
 то >>> 0.01296938731075726 time(s)
 _и >>> 0.0102350746810271 time(s)
 ко >>> 0.010193343300480455 time(s)
 ст >>> 0.010069963566690375 time(s)
 ь_ >>> 0.01004093304109271 time(s)
 по >>> 0.009785101534263279 time(s)
 но >>> 0.009652649761223927 time(s)
 _к >>> 0.009612732788527137 time(s)
 на >>> 0.009307912269751644 time(s)
 _о >>> 0.009258923257805583 time(s)
 я_ >>> 0.00913010030046594 time(s)
 не >>> 0.009033936684423672 time(s)

Ентропія біграми з пробілом 1 крок >>> 3.97403625359661

$$R = 1 - \frac{3.974}{5} = 0.2052 \quad 20.52\%$$

Біграма з пробілом 2 кроки (перші 20)

о_ >>> 0.021333807498584764 time(s)
 и_ >>> 0.018234798891033924 time(s)
 е_ >>> 0.01815496494564034 time(s)
 _п >>> 0.017182442338118532 time(s)
 _н >>> 0.016333299464386802 time(s)
 _в >>> 0.015560361720348947 time(s)
 _с >>> 0.015404322645261492 time(s)

а_ >>> 0.014489861088935015 time(s)
 то >>> 0.01295124323225872 time(s)
 _и >>> 0.010334867112769076 time(s)
 ко >>> 0.010193343300480455 time(s)
 _ь >>> 0.009917553307302629 time(s)
 по >>> 0.009866749887506713 time(s)
 ст >>> 0.009866749887506713 time(s)
 но >>> 0.009779658310713715 time(s)
 _к >>> 0.009641763314124802 time(s)
 на >>> 0.009413147925043183 time(s)
 _о >>> 0.009166388457463022 time(s)
 я_ >>> 0.009075668064970316 time(s)
 не >>> 0.009035751092273525 time(s)

Ентропія біграми з пробілом 2 кроки >>> 3.9730008821084835

$$R = 1 - \frac{3.973}{5} = 0.2054 \quad 20.54\%$$

Біграма без пробілу 1 крок (перші 20)

то >>> 0.015845173234937244 time(s)
 ко >>> 0.012395398281384841 time(s)
 ст >>> 0.01231104641731679 time(s)
 ов >>> 0.012161808503965619 time(s)
 но >>> 0.011796283759670725 time(s)
 по >>> 0.011670837397723365 time(s)
 на >>> 0.01117986629148111 time(s)
 не >>> 0.010842458835208901 time(s)
 ка >>> 0.01019143547406829 time(s)
 ен >>> 0.009975148643124565 time(s)
 ни >>> 0.00939982567281426 time(s)
 го >>> 0.009343591096768892 time(s)
 ак >>> 0.009198678920036596 time(s)
 ос >>> 0.009146770080610102 time(s)
 он >>> 0.009088372636255296 time(s)
 ро >>> 0.008930483249666377 time(s)
 от >>> 0.008854782858836074 time(s)
 ал >>> 0.008824502702503952 time(s)
 пр >>> 0.008539004085658237 time(s)
 ли >>> 0.008247016863884208 time(s)

Ентропія біграми без пробілу 1 крок >>> 4.143481201154285

$$R = 1 - \frac{4.143}{5} = 0.1714 \quad 17.14\%$$

Біграма без пробілу 2 кроки (перші 20)

то >>> 0.015905699145668865 time(s)
 ст >>> 0.012821455607223965 time(s)
 ов >>> 0.012332648426516709 time(s)

```

ко >>> 0.012233156699470098 time(s)
но >>> 0.011982264518222126 time(s)
по >>> 0.011489131610251974 time(s)
на >>> 0.011099816156591327 time(s)
не >>> 0.010840272520817562 time(s)
ка >>> 0.01010489888612523 time(s)
ен >>> 0.009892938250243323 time(s)
ак >>> 0.009529577160160053 time(s)
ни >>> 0.009473342705742403 time(s)
го >>> 0.009443062614902131 time(s)
ос >>> 0.009066724343030172 time(s)
от >>> 0.009053747161241485 time(s)
ро >>> 0.008962906888720666 time(s)
он >>> 0.00887206661619985 time(s)
ал >>> 0.00871634043473559 time(s)
пр >>> 0.008586568616848709 time(s)
во >>> 0.007993943981831945 time(s)

```

Ентропія біграми без пробілу 2 кроки >>> 4.142553385816108

$$R = 1 - \frac{4.142}{5} = 0.1716 \quad 17.16\%$$

➤ CoolPinkProgram

Лабораторная работа №1

Произвольная часть текста:
войне_например_оказалось_бы_лишенным_смысла_какой_смысл_заявлять_что_враг_

Использованные буквы:
_, ш, г, ц, к, н, м, и, ы, й, б, ю, т, с, ч, я, х, з,

Порядок n-граммы:

5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ: п

Символ по счету: 19

Номер эксперимента: 50

Поле ввода символов:
п

Продолжить Другой

Неравенство для энтропии:
 $2,29563319894953 < H < 3,02791993994523$

Двоичная таблица угаданных символов:
01000000000000000000000000000000 ^
00000000000001000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
00100000000000000000000000000000 v

Вероятности:

$q[1] = 0,46$
 $q[2] = 0,08$
 $q[3] = 0,06$
 $q[4] = 0,02$
 $q[5] = 0,08$
 $q[6] = 0$
 $q[7] = 0$
 $q[8] = 0,02$
 $q[9] = 0,04$
 $q[10] = 0$
 $q[11] = 0,02$
 $q[12] = 0$
 $q[13] = 0,02$
 $q[14] = 0,04$
 $q[15] = 0$
 $q[16] = 0$
 $q[17] = 0,02$
 $q[18] = 0,04$
 $q[19] = 0,02$
 $q[20] = 0$
 $q[21] = 0$
 $q[22] = 0$
 $q[23] = 0$
 $q[24] = 0$
 $q[25] = 0,02$
 $q[26] = 0$
 $q[27] = 0,02$
 $q[28] = 0$
 $q[29] = 0,02$
 $q[30] = 0,02$
 $q[31] = 0$
 $q[32] = 0$

Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

$$2.295 < H < 3.027$$

$$H^{(10)} = \frac{2.295 + 3.027}{2} = 2.661$$

Лабораторная работа №1

Произвольная часть текста:
оводу_правил_игры_в_футбол_этот_закон_раньше_называли_естественным_то_есть_

Использованные буквы:

Порядок n-граммы:
 5 символов
 10 символов
 15 символов
 20 символов
 25 символов
 30 символов
 35 символов
 40 символов
 45 символов
 50 символов

Введенный символ: _ (пробел)
 Символ по счету: 1
 Номер эксперимента: 50

Неравенство для энтропии:
 $1,84426142536129 < H < 2,57701481125244$

Двоичная таблица угаданных символов:

01000000000000000000000000000000
00100000000000000000000000000000
10000000000000000000000000000000
00001000000000000000000000000000
10000000000000000000000000000000

Вероятности:

q[1] = 0,48
q[2] = 0,14
q[3] = 0,04
q[4] = 0,14
q[5] = 0,02
q[6] = 0
q[7] = 0,02
q[8] = 0,02
q[9] = 0,02
q[10] = 0
q[11] = 0,02
q[12] = 0,04
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0
q[19] = 0
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0,02
q[26] = 0,02
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0,02

Поле ввода символов:

Продолжить Другой

Строка состояния:
 Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

$$1.844 < H < 2.577$$

$$H^{(20)} = \frac{1.844 + 2.577}{2} = 2.2105$$

Произвольная часть текста:		Вероятности:	
т_же_человек_сам_возвращается_к_отвергнутым_им_принципам_он_может_нарушить_			
Использованные буквы:			
Порядок n-граммы:	Введенный символ: _ (пробел)	Неравенство для энтропии:	
5 символов	Символ по счету: 1	$1,54357349578227 < H < 2,38360999429113$	
10 символов	Номер эксперимента: 50	Двоичная таблица угаданных символов:	
15 символов		00001000000000000000000000000000 ^ 10000000000000000000000000000000 0000000000000000000010000000000000 10000000000000000000000000000000 10000000000000000000000000000000 v ~~~~~~	
20 символов	Поле ввода символов:		
25 символов	<input type="button" value="Продолжить"/>	<input type="button" value="Другой"/>	
30 символов			
35 символов			
40 символов			
45 символов			
50 символов			
Строка состояния:			
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка			

$$H^{(30)} = \frac{1.543 + 2.383}{2} = 1.963$$

➤ Оцінка Надлишковості

$$R_{10} = 1 - \frac{2.661}{5} = 0.4678 \quad 46.78\%$$

$$R_{20} = 1 - \frac{2.2105}{5} = 0.5579 \quad 55.79\%$$

$$R_{30} = 1 - \frac{1.963}{5} = 0.6074 \quad 60.74\%$$

5. Висновки.

Отримав навички роботи з мовою програмування Python. Дізнався як обчислити ентропію, частотність та надлишковість. Розібрався з дослідженням в цілому. Під час дослідження роботи було визначено що найвживаніші букви в російському алфавіті – '_', 'o'; 'e'; 'a'; щодо менш вживаних – 'ё'; 'ъ'; 'ф'; 'э'. Найвживаніші біграми (з пробілом) в російській мові – 'o_'; 'и_'; 'е_'; '_п'; менш вживані (без пробілу) – 'пз'; 'фф'; 'фы'; 'бф'. Також, використавши програму CoolPinkProgram, встановив що середня надлишковість російської мови становить 54.43%.